2006

American sign language - Sentence reproduction test: Development & implications

Peter C. Hauser
Rochester Institute of Technology

Raylene Paludneviciene
Gallaudet University

Ted Supalla
University of Rochester

Daphne Bavelier
University of Rochester

Sign Languages: spinning and unraveling the past, present and future. TISLR9, forty five papers and three posters from the 9th. Theoretical Issues in Sign Language Research Conference, Florianopolis, Brazil, December 2006. (2008) R. M. de Quadros (ed.). Editora Arara Azul. Petrópolis/RJ. Brazil. http://www.editora-arara-azul.com.br/EstudosSurdos.php.

American Sign Language – Sentence Reproduction Test: Development & Implications

Peter C. Hauser
National Technical Institute for the Deaf
Rochester Institute of Technology

Raylene Paludnevičienė
Gallaudet University

Ted Supalla
University of Rochester

Daphne Bavelier
University of Rochester

Corresponding Author: Peter C. Hauser, Department of Research and Teacher Education, National Technical Institute for the Deaf, Rochester Institute of Technology, Rochester, NY, USA 14623; e-mail: peter.hauser@rit.edu

Abstract

The deaf community is widely heterogeneous in its language background. Widespread variation in fluency exists even among users of American Sign Language (ASL), the natural gestural language used by deaf people in North America. This variability is a source of unwanted "noise" in many psycholinguistic and pedagogical studies. Our aim is to develop a quantitative test of ASL fluency to allow researchers to measure and make use of this variability. We present a new test paradigm for assessing ASL fluency modeled after the *Speaking Grammar Subtest* of the *Test of Adolescent and Adult Language,* 3rd Edition (TOAL3; Hammill, Brown, Larsen, & Wiederholt, 1994). The American Sign Language—Sentence Reproduction Test (ASL-SRT) requires participants to watch computer-displayed video clips of a native signer signing sentences of increasing length and complexity. After viewing each sentence, the participant has to sign back the sentence just viewed. We review the development of appropriate test sentences, rating procedures and inter-rater reliability, and show how our preliminary version of the test already distinguishes between hearing and deaf users of ASL, as well as native and non-native users.

1. American Sign Language - Sentence Reproduction Test Development, & Implications

American Sign Language (ASL) is the natural language used by deaf people in the United States and parts of Canada, yet there are tremendous variations in fluency among users of ASL. The need for a robust test of ASL fluency for widespread use is probably felt most pressingly in two main domains. First, proper educational and clinical assessment requires us to evaluate the level of language skill achieved by the individual. However, there is no commonly accepted test to measure language performance in the sign modality. Second, studies of ASL processing and its functional brain organization can only be informative to the extent that they focus on individuals who indeed have native proficiency in the language. In this literature, the absence of an agreed upon benchmark renders participant selection a constant challenge. The lack of precise measurement of the ASL skill of participants is likely to be a source of unwanted noise and, in some cases, may corrupt the results, weakening the validity of the claims made by this body of work.

Research efforts toward a test of ASL proficiency are not a new development (see Haug, 2007 and Singleton, 2003 for reviews). Over the years, several groups have developed tests of ASL skill for use in their research (i.e., Anderson & Reilly, 2002; Boudrealt & Mayberry, 2000; Caccamise & Newell, 1999; Hoffmeister, 1999; Maller. Singleton, Supalla, & Wix, 1999; Mounty, 1994; Shick, 1997; Prinz, Strong, & Kuntze, 1994; Supalla et al., 1999). In general, when reviewing available tests of ASL proficiency, two factors seem to have hampered their wider distribution: lengthy administration time and advanced skill requirements for scoring. A number of these tests were designed to provide a detailed picture of language use in signers, while others give a general profile of language proficiency. This is the case, for example, with the checklist paradigm developed by Maller and colleagues (1999) for deaf children aged 6 – 12. Their test, the *American Sign Language Proficiency Assessment*, covers a reasonable range of linguistic skills as well as interpersonal interaction skills in open-ended dialogue situations. Although the test takes approximately 30 minutes to collect specific sign samples, it takes 1– 2 hours for skilled raters to code each participant's sign samples.

The ASL proficiency tests that provide detailed assessment of several different linguistic constructs elicited under controlled situations are extremely valuable for furthering our understanding of ASL processing. However these tests generally are time consuming to administer and require significant expertise and familiarity with the linguistic constructs being measured in order for them to be to be scored adequately. Many of the existing proficiency tests take one hour or more to administer and have components that range from one hour (Prinz, Strong, & Kuntze, 1994) to 15 or more hours to score

(Hoffmeister, 1999; Supalla et al.1999). ASL tests that take less than one hour to administer do exist, but still have time-consuming scoring procedures that take one hour or more to complete (Caccamise & Newell, 1999; Maller et al., 1999). These tests require elicitors and/or interviewers, which is a problem, as most laboratories do not have such trained or skilled personnel. In addition, qualitative assessment by individuals prevents direct comparison across groups and studies. Some tests may be quickly administered and scored but provide assessment of specific linguistic constructs rather than global ASL fluency (Anderson & Reilly, 2002; Boudreault & Mayberry, 2000; Schick, 1997).

Typically, ASL proficiency tests focus either on children (Anderson & Reilly, 2002; Hoffmeister, 1999; Maller et al., 1999; Schick, 1997; Prinz, Strong, & Kuntze, 1994) or on adults (Caccamise & Newell, 1999). Only one test is adequate for use with both children and adults: the Test Battery for ASL Morphology and Syntax (Supalla et al., 1999), however, it takes approximately two hours to administer and 15 hours to score. There remains therefore a need for a quantitative test of global ASL proficiency for both children and adults that involves a short administration time and scoring procedures that are robust across scorers and can be completed by individuals without intensive training.

We present here the development stages of a new test of ASL proficiency modeled after the *Speaking Grammar Subtest* of the *Test of Adolescent and Adult Language—Third Edition* (TOAL3; Hammill, Brown, Larsen, & Wiederholt, 1994). In the Speaking Fluency Subtest, participants listen to English sentences of increasing length and complexity. After each sentence, they are asked to repeat verbatim the sentence they just heard. Any departure from verbatim recall is considered an error. This test was selected as a model for our project because performance on this test reliably identifies language impairment in children and young adults (Hammill et al., 1994); in addition, it has been shown to distinguish native speakers of English from late-learners (Newman, Waligura, Neville, & Ullman, 2003).

Our aim in adapting this test into what we will term the American Sign Language—Sentence Reproduction Test (ASL-SRT) is to converge on a test that is sensitive enough to discriminate between individuals who acquired ASL from their parents beginning at birth and those who did not (thereafter referred to as native signers and non-native signers, respectively). In addition, the test must discriminate within the native signer population individuals fully mastering ASL versus those not yet achieving full mastery.

This paper describes the development of the ASL test sentences, and piloting on populations with varied ASL background. In the first phase of test development, it was necessary to test the items' difficulty and effectiveness in separating different levels of ASL proficiency. The results of these analyses are presented here. This data will be used to determine the items to be included in the final version of the ASL-SRT and their arrangement in order of increasing difficulty.

2. Method

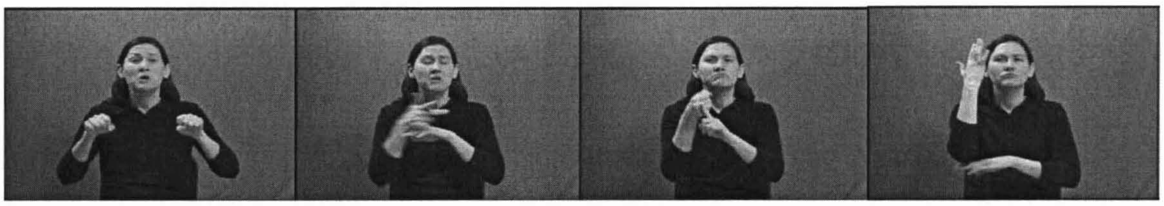
In the Speaking Grammar Subtest of TOAL3, participants are required to repeat orally presented sentences verbatim. Thirty English sentences are administered in increasing order of difficulty. One point is awarded for each sentence that is reproduced correctly and zero points are awarded for sentences that are reproduced with omissions or commissions. An aspect of the administration procedure that was not used in the ASL-SRT was the fact that the Speaking Grammar Subtest is usually administered live with the examiner tracking the performance of the subject as testing proceeds. Because our aim was (1) to develop a test that does not required a skilled interviewer and (2) to collect a full range of data from widely diverse groups of ASL users to determine complexity ranking of sentences, our ASL version presented pre-recorded videos of a native signer (author RP) signing all test sentences.

Forty sentences that increase in length and syntactic, thematic and morphemic complexity were newly developed for the ASL-SRT (see Figure 1 for examples). There are several issues specific to ASL that have to be taken into consideration when developing such a set of sentences. The first issue concerns lexical item selection. Our goal was to have a set of sentences that could be reproduced by ASL signers from different regions and of different ages, with a good distribution of difficulty to avoid floor or ceiling effects. Thus, when identifying signs for use in sentences, regional signs, signs that vary from generation to generation, and sign system variations were avoided. Sentence complexity was increased by allowing fingerspelling, signs with numerical incorporation affixes, and signs with assumed low frequency of occurrence in the language. The test sentences were also built so as to span a wide range from very basic to quite complex, given the widespread variations in sign proficiency observed in sign users.

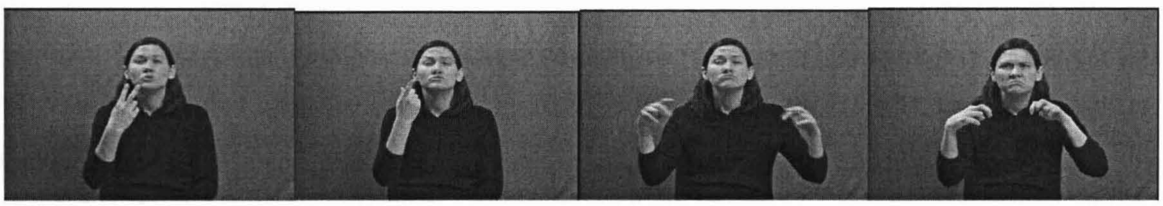
A second issue is that in signed languages, longer sentences do not necessarily translate to more difficult sentences. Instead, a shorter sentence may be morphologically more complex than a longer sentence due to the way grammatical features are expressed in individual signs (e.g., polymorphemic signs). For example, classifier morphemes are used in descriptive constructions comprised of multi-layered tiers of handshapes, locations and movements. Such forms increase the difficulty level of a sentence without making it longer. For example, the sentence "MOTORCYCLE SPIN HIT TREE" can be roughly translated by only four English glosses that describe the general concept of each sign (see Figure 1). However, the English gloss for SPIN does not convey all of the concepts expressed by the sign. SPIN was signed with the 3 handshape, which is a classifier indicating a vehicle. The base hand faces downward with the palm open representing the surface that the vehicle is on. The movement of the sign shows the vehicle moving from back to side. This single signed form conveys a wealth of information that is missing from the English gloss (and translation in the figure), and would require a higher level of cognitive processing, likely corresponding to a more difficult sentence.

Taking these factors into consideration, the sentences were arranged in the order of increasing complexity as determined by two deaf native signers (co-authors Paludnevičienė, Supalla). The sentences were pilot tested on a group of eight signers with varying skills in ASL, including novice hearing adult signers, hearing native and non-native ASL interpreters, and deaf native and non-native adult signers. Based on the results of this pilot study, sentences were re-ordered and changed as needed. In particular, sentences were modified to create a more even distribution of easy to difficult sentences and to avoid the use of grammatical or lexical features that were found to vary in the pilot participants' sentence reproductions. A refined set of 39 sentences was further tested on a large number of participants to determine: (1) how well this material may differentiate native and non-native signers, deaf and hearing individuals, and children and adult signers; (2) whether rating of relative difficulty could be reliably derived by different scorers; and, (3) the extent to which a final sentence order could be agreed upon. The data from this study are presented here and will be used to create the final version of the ASL-SRT.

Easy Sentence: "The motorcycle spun and hit a tree."



Difficult Sentence: "Two dogs walking by noticed each other, started growling, then jumped into a fight while people rushed in to pull them apart."



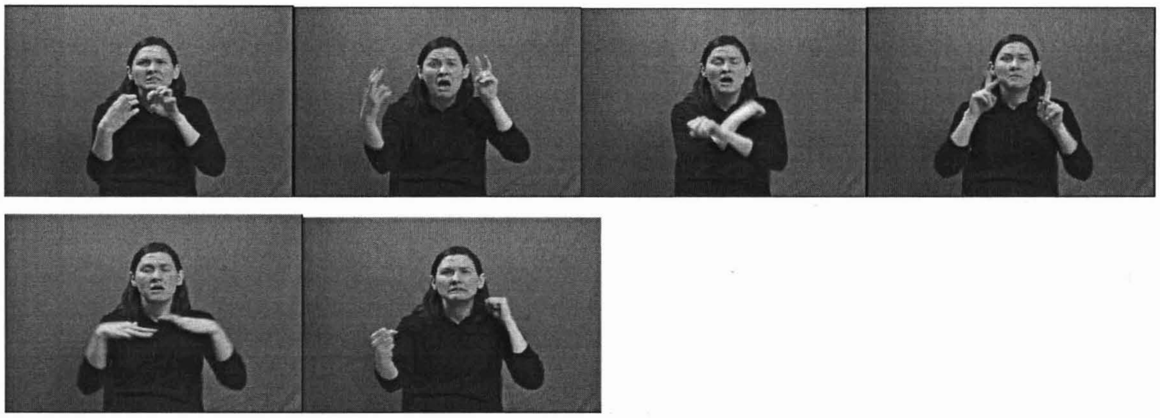


Figure 1. ASL-SRT examples of an easy sentence (top) and a difficult sentence (bottom). English translations are provided instead of sign glosses.

2.1. Participants

The second phase of the ASL-SRT was administered to 120 deaf and hearing children (M_{age} = 12.9, SD_{age} = 1.7) and adults (M_{age} = 27.6, SD_{age} = 10.1) who are native or non-native signers. The participants were recruited from summer camps, schools, and colleges for the deaf and grew up in different regions of the United States. The participants had varying proficiency in ASL, ages, and hearing status (see Table 1). Data from six participants were excluded from further analyses due to technical difficulties during administration or confusion regarding test instructions.

	Deaf		Hearing	
	n	M_{age} (*SD*)	*n*	M_{age} (*SD*)
Children				
Non-native	13	14.1 (1.8)	0	-
Native	27	12.5 (1.5)	5	11.8 (1.9)
Adult				
Non-native	4	19.3 (1.9)	2	40.0 (14.1)
Native	23	23.4 (4.4)	25	31.8 (11.7)

Table 1. Number and age of deaf and hearing native and non-native children and adult signers.

2.2. Procedure

The test and its instructions were entirely administered through pre-recorded videos featuring a deaf native ASL user (co-author Paludnevičienė). Participants were tested individually. After viewing the instruction video, participants were then allowed to ask clarification questions to the investigator. To ensure that the participants understood the instructions, two practice sentences were given. After viewing each practice sentence, participants were asked to reproduce it. Participants were not corrected for the mistakes they made, but reminded that the goal of the task was to reproduce the sentence as signed by the model and not just reproduce its gist. The videos of the ASL-SRT sentences were then presented to participants, one sentence at a time in order of increasing complexity. Participants' responses were recorded on video and they pressed the space bar each time they were ready for the next sentence.

In this preliminary version of the ASL-SRT, all 39 sentences were administered independently of the participants' performance. Two deaf native ASL signers rated the participants' responses off line. The raters were undergraduate psychology students with no formal education in linguistics and only brief training on how to rate the video data. They were trained to watch the original sentence on video and then watch the participants' reproduction of the sentence. Scorers were instructed to mark a sentence as incorrect if they felt that the reproduction was not exactly the same as the original. The raters were also asked to describe errors in detail on a scoring sheet. While this is not intended to be a part of the final scoring procedures of the ASL-SRT, this data was requested for analyses of discrepancies between raters, and as a source of examples for the development of a rater manual for the final version of the test illustrating errors and acceptable variations in signs.

Sample sentence	**MOTORCYCLE SPIN HIT TREE**
Error Types	
omissions	MOTORCYCLE HIT TREE
comissions	MOTORCYCLE SPIN HIT TREE FALL
syntax	MOTORCYCLE SPIN TREE HIT
variation	MOTORCYCLE SPIN (base hand up) HIT TREE
comprehension	MOTORCYCLE RIDE SAW TREE

Figure 2. Sample of ASL-SRT error types.

One of the co-authors (Paludnevičienė) reviewed the raters' scores on a single subject, provided feedback and answered their questions. No further instructions were provided to the raters after their training, as our goal was to develop a test that required minimal training, and the two raters did not communicate about their ratings or progress with each other. The scoring system awarded one point for each sentence that was signed correctly. No points were awarded to sentences that had omissions, commissions or changes in lexical items or syntactic constructions (see Figure 2 for examples of error types).

A total of 99 of the participants were scored by both raters and used in the following analyses. Data from the 99 participants were used to calculate reliability analyses. Data from the 48 hearing and deaf native signers were used to test the effect of hearing status on ASL-SRT scores. Data from the 67 younger deaf participants were used to determine the effect of age and native exposure on the ASL-SRT scores (see Table 1 for n and M_{age} for each group).

3. Results

3.1. Reliability

The mean scores for each of the raters were first computed. A paired-samples t-test revealed that grand mean scores from Rater 1 (M = 24.6, SD = 6.6) and Rater 2 (M = 18.8, SD = 6.3) were significantly different, $t(98) = 15.28$, $p < .001$. One rater had a higher criterion for accepting sentence reproductions as correct compared to the second rater. Despite this difference in baseline, significant inter-rater reliability was still found, Pearson $R = .83$, $p < .01$. This high reliability coefficient indicates a significant amount of agreement as to whether raters felt the sentence was correctly or incorrectly reproduced by the participants. In addition high internal consistency was found within each rater's score, alpha coefficients are .87 (Rater 1) and .88 (Rater 2). The participants' scores obtained from the two raters revealed high inter- and intra-rater reliability. The mean of the two rater's scores for each sentence was combined for the following analyses.

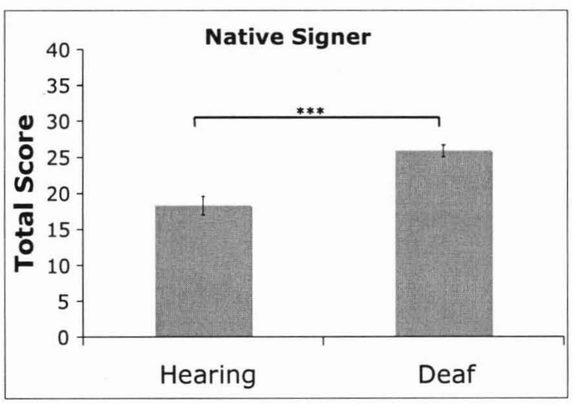


Figure 3. Deaf and hearing adult native signers' ASL-SRT performance.

3.2. Effect of Hearing Status in Adults

To assess whether deaf and hearing signers could both be used in the development of the sentence rank order required by the test, we first compared deaf and hearing native signers' overall performance on all 39 sentences included in the test. Our analysis focused on adult signers. However, we experienced difficulty with screening control and finding hearing signers who had undergone a level of immersion in the language that was similar to deaf native signers. Even highly fluent hearing interpreters and professionals are more likely to achieve mastery via formal instruction rather than immersion. An independent sample *t*-test revealed that the mean total correct sentence reproductions made by the 25 hearing native signers (M = 18.3, SD = 6.3) was significantly worse compared to the performance of the 23 deaf native signers (M = 25.9, SD = 4.0), $t(46) = -4.95$, $p < .001$). The results of this analysis suggest that hearing and deaf adults have different levels of proficiency in ASL (see Figure 3). Therefore, the data from hearing participants will not be used to rank the sentences in order of difficulty for the final test product.

	Native		Non-native	
	n	Mean (*SD*)	*n*	Mean (*SD*)
Deaf Children	27	23.9 (4.4)	13	17.1 (7.8)
Deaf Adults	23	25.9 (4.3)	4	22.2 (5.4)

Table 2. Deaf children and adult native and non-native signers' ASL-SRT scores

3.3. Effect of Native Fluency and Developmental Age in Deaf

We relied on the performance of deaf signers to determine sentence rank order. To insure enough variations in skill level, we have included native and non-native signing children between the ages of 10 to 17 years old, and adults between the ages of 18 and 60 years old. Before turning to sentence ranking, it was necessary to first confirm that our set of sentences as a whole was sensitive to developmental age and native fluency. A 2x2 ANOVA with native fluency (native vs. non-native) and developmental age (children vs. adults) was computed using the total scores on ASL-SRT (see Table 2). The results of the analysis revealed a main effect of native fluency (M_{native} = 24.8, SD_{native} = 4.1; $M_{non\text{-}native}$ = 18.3, $SD_{non\text{-}native}$ = 7.2, $F(1, 63) = 11.33$, $p = .001$), and of developmental age ($M_{children}$ = 21.7, $SD_{children}$ = 6.2; M_{adults} = 25.3, SD_{adults} = 4.2, $F(1, 63) = 5.10$, $p < .05$) without an interaction effect. The results reveal that the ASL-SRT is sensitive to differences in ASL fluency between native and non-native signers. It also distinguishes between children and adults (see Figure 4).

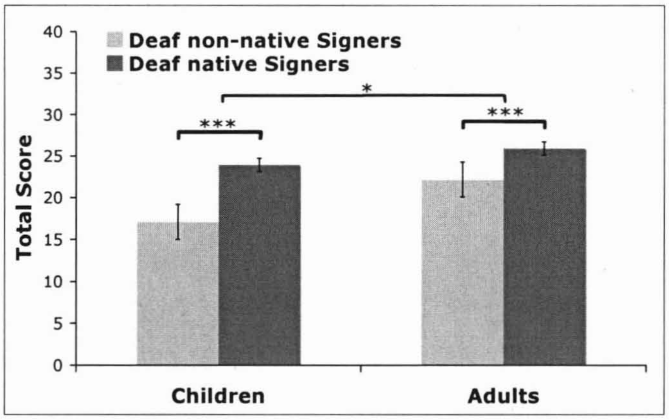


Figure 4. Deaf adult and children native and non-native signers' ASL-SRT scores.

3.4. Item Reordering

The final goal of the ASL-SRT development is to rank order sentences according to their difficulty level, as being done in the TOAL3. We used only the data from the children native signers to determine item difficulty so that the test can begin with the easiest sentences and end with the most difficult ones. One motivating factor for not including deaf non-native signers is the possible discrepancy in developmental pattern. The psycholinguistic and sociolinguistic literature (e.g., Newport & Meier, 1985; Mayberry, 1994; Ross & Newport, 1996) has generally shown that deaf non-native signers appear to be less fluent than deaf native signers. Deaf non-native adult signers, although providing a wide range of skill level, may fail on some sentences because of their lack of native exposure rather than the true linguistic difficulty of the items. This discrepancy raises an issue as to whether the mistakes that non-native signers make are representative of the native user population, and for that reason non-native signers were ruled out. Deaf children who are native signers were the best candidates as their performance reflects native usage and yet they still exhibit a wide range of skill levels providing the necessary information to rank order the sentences.

The two raters' mean total score of each item from the children native ASL signers' data was used to establish the new ASL-SRT item order that is shown as a prototype model with 39 test sentences in Figure 5 on the right. Items were reordered from the items with the highest percentage correct to lowest. The two graphs on the left of the figure illustrate the total percentage correct for each item as scored by the two raters. Out of these 39, we will select a final set of 30 sentences that will be then checked again on a new sample of deaf signers.

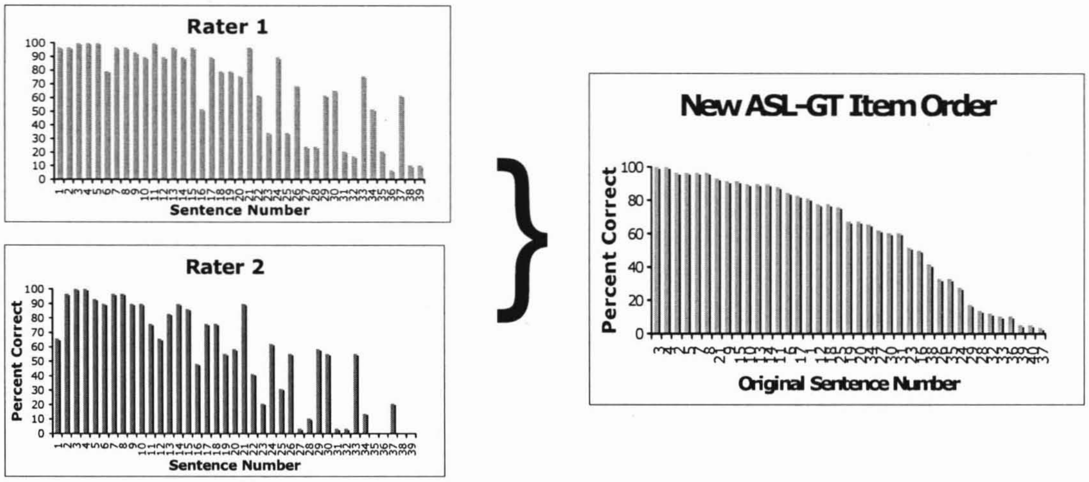


Figure 5. Reordering of the ASL-SRT items based on the deaf children native signers' data.

4. Discussion

The results presented here represent the first attempt in ranking test sentences for the ASL-SRT. One compelling reason for pursuing this route for measuring ASL proficiency is that the test is easy to administer and can be scored with robust inter-rater reliability and internal consistency by native signers with no linguistic background and minimal training. In its present format, it seems sensitive to differences between native and non-native ASL signers and captures the difference between children and adult deaf signers. However, the claims made here about the effects of hearing status, and developmental age of exposure on native ASL fluency need to be confirmed with a wider population. It will also be necessary to demonstrate the inter-rater reliability and internal consistency of the final version.

Ideally, the final version of the ASL-SRT will be optimized by including only the items that best predict native fluency in ASL and are most sensitive to age differences in language skills. Qualitative analyses of participant errors are currently being performed to determine whether there are some natural sign variations that should be permitted (e.g., weak hand drop). The analyses will also help the authors explore how the raters determined whether sign reproductions were correct or incorrect. This information will be used to develop a rater training manual for the final version of the test.

Validity testing will also be necessary to confirm that the ASL-SRT actually measures ASL competence. One way to illustrate a test's validity is by investigating whether the test can discriminate between different groups of signers (discriminant validity) using similar methods as presented here. It will also be necessary to determine the test's concurrent validity by demonstrating that the participants' ASL-SRT scores positively correlate with another measure of ASL competence. Finally, separate normative samples will be collected from child and adult native and non-native deaf ASL signers. The final set of sentences will be tested on a large sample of deaf and hearing individuals to confirm the results described here.

The ASL-SRT is a global, objective ASL proficiency test that involves receptive and expressive skills in both children and adults. This test takes less than 30 minutes to administer and less than 30 minutes to score by deaf native signers with minimal training. Currently, the inter-rater reliability and internal consistency of the preliminary version of the ASL-SRT are already high. To shorten the administration time, we will reduce the number of items in the final version. We will also develop an administration and scoring manual to establish standardization, and will include specific rater training components that will hopefully also shorten the scoring time. Upon completion of the development project and psychometric studies, the ASL-SRT test will be available for researchers' use for cross-study comparisons.

References

Anderson, D. & Reilly, J. (2002). The MacArthur Communicative Development Inventory: Normative Data for American Sign Language. *Journal of Deaf Studies and Deaf Education, 7, 2, pp. 83-119.*

Boudreault, P., & Mayberry, R. (2006). Grammatical processing in American Sign Language: Age of first-language acquisition effects in relation to syntactic structure. *Language and Cognitive Processes, 21,* 608-635.

Caccamise, F. & Newell, W. (1995). Evaluating Sign Language Communication Skills: The Sign Communication Proficiency Interview (SCPI). In R. Myers (Ed.), *Standards of Care for the Delivery of Mental Health Services to Deaf and Hard of Hearing Persons* (pp. 33-35). Silver Spring, MD: National Association of the Deaf.

Hammill, D., Brown, V., Larsen, S., & Wiederholt, J.L. (1994). *Test of Adolescent and Adult Language, Third Edition.* Austin, Texas: PRO-ED, Inc.

Haug, T., (2007). *Sign Language Assessment.* Retrieved March 15, 2007, from http://www.signlang-assessment.info/

Hoffmeister, R. J. (1999). *American Sign Language Assessment Instrument (ASLAI).* Manuscript, Center for the Study of Communication & the Deaf, Boston University.

Maller, S. J., Singleton, J. L., Supalla, S. J. & Wix, T. (1999). The Development and Psychometric Properties of the American Sign Language Proficiency Assessment (ASL-PA). *Journal of Deaf Studies and Deaf Education, 4,* 249-269.

Mayberry, R. I. (1994). The importance of childhood to language acquisition: Insights from American Sign Language. In J. C. Goodman & H. C. Nusbaum (Eds.), *The Development of Speech Perception: The Transition from Speech Sounds to Words* (pp. 57-90). Cambridge: MIT Press.

Mounty, J. (1994). *Signed Language Development Checklist.* Princeton, NJ: Educational Testing Service.

Newman, A.J., Waligura, D.L., Neville, H.J., & Ullman, M.T. (2003). Effects of late second language acquisition on neural organization: Event-related potential and functional magnetic resonance imaging evidence. *Cognitive Neuroscience Society Abstracts, 11.*

Newport, E. & Meier. R. (1985). The acquisition of American Sign Language. In D. Slobin (Ed.), *The crosslinguistic study of language acquisition*: Vol 1, (pp. 881-938). Hillsdale, NJ: Erlbaum.

Prinz, P., Strong, M. & Kuntze, M. (1994). *The Test of American Sign Language*. Unpublished test. San Francisco State University, California Research Institute.

Ross, D. S., & Newport, E. L. (1996).0 The development of language from non-native linguistic input. In A. Stringfellow, D. Cahana-Amitay, E. Hughes, & A. Zukowski (Eds.), *Proceedings of the 20th Annual Boston University Conference on Language Development*, vol. 2, 634-645.

Singleton, J. (2003). Assessing childrens's proficiency in natural signed languages. In M. Marschark & P. E. Spencer (Eds.), *Oxford Handbook of Deaf Studies, Language, and Education* (pp. 289 – 304). New York: Oxford University Press.

Supalla, T., Newport, E., Singleton, J., Supalla, S., Coulter, G. & Metlay, D. (1995). *An Overview of the Test Battery for American Sign Language Morphology and Syntax*. Paper presented at the Annual Meeting of the American Educational Research Association (AERA), April 20, 1995, San Francisco, CA.

Acknowledgements

This research was supported by NIDCD grants to D.B. (DC04418) and T. S (DC00167). We would like to express our appreciation for the support and participation of all people in this study. Special thanks to Dara Baril, Matt W.G. Dye, Matt Hall, Wyatte Hall, Tiffany Panko, Alex Pouliot, Betsy Hicks McDonald, Carissa Thompson and campers and staff at Camp Lakodia and Camp Mark 7. We are greatly indebted to Aaron Newman and Elissa Newport for their critical and constructive feedback on this project.